

Faculty of Engineering and Information Technology  
University of Technology Sydney

# **High-density Visual Crowd Counting with Perspective Understanding in Deep Neural Networks**

A thesis submitted in partial fulfillment of  
the requirements for the degree of  
**Doctor of Philosophy**

by

Muming Zhao

January 2020

## CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

This thesis is the result of me conducted jointly with Shanghai Jiao Tong University as part of a collaborative Doctoral degree.

Signature of Candidate: Production Note:  
Signature removed  
prior to publication.

Date: 2020/01/23

# Acknowledgments

I would like to thank my principal advisor Jian Zhang in UTS, for his continuous support either in my research or in the finance. He had always expressed his patience to me and never gave me up even at the time when I stuck in my progress. His enthusiasm and keen sense to the research helped guide and push me to complete different works in this thesis. I also want to express my gratitude to my co-supervisor Chongyang Zhang in SJTU for his significant support of my research. He used to spend a lot of time to help me revise my paper and give me valuable comments to improve my academic skills. I also would like to thank my principal supervisor Wenjun Zhang in SJTU, for his continuous support of my dual PhD study in UTS and SJTU. Without the three supervisors, this thesis would be impossible.

I want to thank all my colleagues: Xiaoshui Huang, Yazhou Yao, Junjie Zhang, Jiangchao Yao and Yuangang Pan. I appreciate the time they have spent discussing with me, where I have got inspired a lot.

Finally and most essentially, I am grateful for all the support from my parents, my sister and my dear friends. They are the source of my strength.

# Contents

<b>Certificate</b>	<b>i</b>
<b>Acknowledgment</b>	<b>ii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Publications</b>	<b>xiii</b>
<b>Abstract</b>	<b>xiv</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Background	1
1.2 Scope and Limitation of Current Research	3
1.3 Research Contribution	6
1.4 Thesis Structure	7
<b>Chapter 2 Literature Review</b>	<b>9</b>
2.1 Counting by Detection	9
2.2 Counting by Clustering	12
2.3 Counting by Regression	13
2.3.1 Direct Regression	14
2.3.2 Density-based Regression	16
2.4 Counting with Deep Neural Network	17
2.4.1 Convolutional Neural Network	18
2.4.2 Recurrent Neural Networks and Long Short-term Memory Networks	20
2.4.3 DNN-based Crowd Counting	21

<b>Chapter 3</b>	<b>Scale-aware Crowd Counting via Depth-embedded Convolutional Neural Networks . . . . .</b>	<b>27</b>
3.1	Introduction . . . . .	28
3.2	Approach . . . . .	31
3.2.1	Overview . . . . .	31
3.2.2	Depth Prediction . . . . .	32
3.2.3	Depth Embedding Module . . . . .	34
3.2.4	Depth Embedded Network (DeemNet) . . . . .	38
3.3	Model Training . . . . .	39
3.4	Experiments . . . . .	39
3.4.1	Implementation . . . . .	39
3.4.2	Datasets . . . . .	40
3.4.3	Diagnostics Experiments . . . . .	42
3.4.4	Comparison with State-of-the-art . . . . .	45
3.5	Conclusion . . . . .	49
<b>Chapter 4</b>	<b>Towards Locally Consistent Object Counting with Constrained Multi-stage Convolutional Neural Networks . . . . .</b>	<b>50</b>
4.1	Introduction . . . . .	51
4.2	Relationship Between Global Counting Errors and Local Counting Errors . . . . .	53
4.3	Constrained Multi-stage Convolutional Neural Networks . . . . .	54
4.3.1	Density Map Based Object Counting . . . . .	54
4.3.2	Multi-stage Convolutional Neural Network . . . . .	54
4.3.3	Grid Loss . . . . .	57
4.4	Experimental Results . . . . .	58
4.4.1	Implementation . . . . .	58
4.4.2	Ablation Experiments . . . . .	61
4.4.3	Comparison with the State-Of-The-Arts . . . . .	61
4.5	Conclusions . . . . .	64

<b>Chapter 5</b>	<b>Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting . . . . .</b>	<b>66</b>
5.1	Introduction . . . . .	67
5.2	Methodology . . . . .	69
5.2.1	Auxiliary Tasks Prediction . . . . .	70
5.2.2	Main Tasks Prediction . . . . .	74
5.2.3	Optimization . . . . .	74
5.3	Implementation . . . . .	74
5.4	Experiments . . . . .	75
5.4.1	Datasets . . . . .	75
5.4.2	Diagnostics Experiments . . . . .	76
5.4.3	Comparison with State-of-the-art . . . . .	78
5.4.4	Parameter Study of the Weights for Auxiliary Tasks . . . . .	82
5.5	Conclusion . . . . .	85
<b>Chapter 6</b>	<b>Conclusion and Outlook . . . . .</b>	<b>86</b>
6.1	Conclusion . . . . .	86
6.2	Short-term Outlook . . . . .	88
6.2.1	Semi-supervised and Weakly-supervised Learning . . . . .	88
6.2.2	Model Adaption . . . . .	89
6.2.3	Multi-view Crowd Counting . . . . .	89
<b>Bibliography</b>	<b>. . . . .</b>	<b>91</b>

# List of Figures

1.1	Illustration of crowded scenes. . . . .	2
1.2	Sample of a pair of image and its corresponding ground-truth density map for density-based counting methods. . . . .	4
3.1	Our motivation (best viewed in color): Due to scale changes of pedestrians, the three regions (black, orange and red circles) that occupy the same number of pixels have different crowd counts; 6 in the far field (black), 3 in the midway (orange), and 1 in the near field (red) respectively. Since these three regions have the same area, the density values within the farthest circle should be larger than the ones in the nearer circles. In other words, objects with smaller scales should have larger density values and vice versa. This can be interpreted as <i>scale-aware</i> density values. . . . .	29

3.2	Overview of the proposed Deem-CNN. For the $l$ -th layer in the CNN encoder, initial feature maps $\mathbf{Z}^l$ is the output of the previous $(l-1)$ -th layer. We build a Depth Embedding Module on top, including a depth encoding layer, a depth rectifying layer and a depth embedding layer to capture essential geometric depth cues to predict attentive scale-aware scaling weights $\gamma^l$ that are conditional on the feature maps and the predicted depth result. The learned weights re-calibrate the magnitude of features at individual location, results a weighted scale-aware feature map $\mathbf{X}^l$ . . . . .	31
3.3	Visualization of depth maps from the pre-trained DCNF model for depth prediction (Liu, Shen, Lin & Reid 2016). The first row shows sample images from four crowd counting datasets (Zhang, Li, Wang & Yang 2015, Zhang, Zhou, Chen, Gao & Ma 2016, Idrees, Saleemi, Seibert & Shah 2013, Chen, Loy, Gong & Xiang 2012), respectively. The first three images all depict outdoor scenes while the last one is from an indoor scene. The second row visualizes the predicted depth map of each sample image. . . . .	33
3.4	Visualization of attention masks. The first column shows two sample images. The second and the third column respectively visualizes the learned attention masks when the attention module is set at increasing depths of the backbone model. In all the heat maps from blue to red, the underlying value becomes larger. . . . .	36
3.5	Visualization of the image (first row), attention mask (second row), the depth map shown in color (third row) and the generated attentive scale-aware weight maps after depth rectification (last row). . . . .	37



3.6	Sample images from the four evaluation datasets: ShanghaiTech (Zhang et al. 2016), WorldExpo'2010 (Zhang et al. 2015), UCF_CC_50 (Idrees et al. 2013) and the Mall (Chen et al. 2012).	41
3.7	Qualitative visualization. From the first to the last column are: the images, estimated density maps without using the depth embedding module (CSRNet), estimated density maps with the depth embedding module (Deem-CSRNet) and the ground truth density maps. Crowd counts are labeled on the top, and local counts for each one-quarter-sized sub-regions of the image are also labeled for comparison.	48
4.1	Illustration of a locally inconsistent density map prediction. (a) to (c): the original image, the ground truth and the estimated density map. We observe that although the estimated total count (shown in the upper right box) is very close to the ground truth, the quality of prediction is not satisfactory with observation of obvious background noise and count errors of local regions (shown in the red-line-framed boxes).	51
4.2	Architecture of the multi-stage convolutional neural network. We stack several base models sequentially with feature conversion blocks which i). perform feature dimension alignment of feature maps between two adjacent base models, and ii). generate a prediction for each base model to enable intermediate supervision. The first base model accepts the input image, and the rest base models in the following stages accept feature maps which comes from the previous feature conversion block.	55
4.3	Effects of the grid loss on a three-stage model. It can be observed that training with grid loss drives the model to learn to correct the regression errors and produce more accurate object counting results.	58

4.4	Density map prediction results as input images proceed through the multi-stage convolution model. The first row lists images sampled from the ShanghaiTech dataset (first two) and the TranCos dataset (last one). The second to the fourth rows show the intermediate outputs from the first two stages and the final prediction of the last stage, respectively. The ground truth density maps are shown in the last row. Object count derived from the density map are labeled on top of each prediction result. For the first two crowded sample images we also randomly select several subregions to track the local object counts, which are shown in the red boxes. . . . .	63
5.1	Motivation. . . . .	67
5.2	Overview of the proposed approach with the learning of three auxiliary tasks in CNNs (AT-CNN). The symbols of L1 to L3 denote the losses to optimize the auxiliary tasks of crowd segmentation, depth prediction and count regression. The symbols of L4 is the loss for the main task of density estimation. . . . .	69
5.3	Label generation for auxiliary tasks. Given a pair of crowd image and its ground truth density map (the first column), the depth map can be estimated using external depth prediction algorithms (Liu et al. 2016) and the crowd segment is inferred through binarization of the density map (the second column). The distilled depth map (the third column) used to supervise the auxiliary task is obtained by masking the original estimated depth map with the crowd segment map. . . . .	72
5.4	(a) Histogram: comparison of average count estimation on 10 splits of ShanghaiTech-B dataset according to the increasing number of people in each image. (b) Visualization of a failure case from the last split. . . . .	81

5.5	Comparison of MAE with different weight of the loss for the three auxiliary tasks on ShanghaiTech-B dataset (Zhang et al. 2016)	83
5.6	Visualization and comparison of density estimation. First column: test image; Second column: depth map predicted by the depth decoder of our method; Third Column: crowd segmentation predicted from the segment decoder of our method; Fourth column: estimated density map by CSRNet (Li, Zhang & Chen 2018); Fifth column: estimated density map by our method (At-CSRNet); Last column: Ground-truth density maps. Count estimation from each density map are labeled at the right corner of the corresponding prediction.	84

# List of Tables

3.1	Different encoder-decoder architectures evaluated in the experiment. . . . .	40
3.2	Component analysis on ShanghaiTech-B dataset. In each stage the best MAE/MSE is indicated as <b>bold</b> and the second best as <i>Italic</i> . . . . .	43
3.3	Diagnostic experiments on ShanghaiTech-B dataset on the number of depth embedding modules. Number of $n$ denotes $n$ proposed modules which are respectively added in the first $n$ stage of the base CNN. . . . .	44
3.4	Comparison results of different methods on the ShanghaiTech-B.	45
3.5	Comparison results of MAE on WorldExpo'2010 dataset. . . .	46
3.6	Comparison results of MAE and MSE on UCF_CC_50 dataset.	47
3.7	Comparison results of MAE and MSE on Mall dataset. . . . .	47
4.1	Performance of ablation experiments for network structures and supervisions. . . . .	60
4.2	Comparison results on the ShanghaiTech dataset. . . . .	62
4.3	Comparison results of GAME on the TRANCOS dataset. . . .	64
5.1	Different encoder-decoder architectures evaluated in the experiment. . . . .	76
5.2	Diagnostic experiments of AT-CFCN and AT-CSRNet on the ShanghaiTech-B dataset (Zhang et al. 2016). . . . .	78

5.3	Diagnostic experiments of AT-CFCN on the Mall dataset (Chen et al. 2012). Dep, Seg and Cot represents the corresponding auxiliary task of depth prediction, crowd segmentation and count regression, respectively. . . . .	79
5.4	Comparison with other state-of-the-art crowd counting methods on the ShanghaiTech-B dataset (Zhang et al. 2016). . . .	80
5.5	Comparison with other state-of-the-art crowd counting methods on the Mall dataset (Chen et al. 2012). . . . .	80
5.6	Comparison with other state-of-the-art crowd counting methods on the WorldExpo'2010 dataset (Zhang et al. 2015). . . .	81

# List of Publications

## Papers published

- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Wenjun Zhang: Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting, *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Wenjun Zhang: Towards Locally Consistent Object Counting with Constrained Multi-stage Convolutional Neural Networks, *14th Asian Conference on Computer Vision (ACCV)*, 2018.
- **Muming Zhao**, Jian Zhang, Fatih Porikli, Chongyang Zhang, Wenjun Zhang: Learning a perspective-embedded deconvolution network for crowd counting, *IEEE International Conference on Multimedia and Expo (ICME)* , 2017.

## Papers in submission

- **Muming Zhao**, Jian Zhang, Chongyang Zhang, Fatih Porikli, Bingbing Ni, Wenjun Zhang: Scale-aware Crowd Counting via Depth-embedded Convolutional Neural Networks, *Transactions on Circuits and Systems for Video Technology (TCSVT)*, *under review*.

# Abstract

With population growth and worldwide urbanization, crowd gathering in public places has become more common. Thus estimating the number of people and measuring their density has become essential for practical applications such as physical security control and public space management. However, the complex environments of crowded scenes have imposed several challenges to general counting algorithms, among which scale variations of pedestrians is one of the most significant problems. With varying-sized objects, it is rather difficult for density-based counting systems to generate appropriate density estimations that conform to scale variations, which usually significantly degrades the counting accuracy. To handle the perspective distortion and the related scale-variation problem, traditional methods mainly perform feature normalization for perspective correction. However, within the deep learning framework, the perspective distortion has not been explicitly considered and addressed. Can we extend the mechanism of perspective handling with the powerful deep learning technique for further improvement? In this dissertation, we focus on measuring crowd density through deep architectures with in-network perspective understanding. Three works are presented. First, we develop a depth-embedded network that augments the original features to be scale-aware for more accurate density estimation. The depth map of a scene is encoded, rectified and finally embedded into the network via a proposed depth embedding module. Thus the objects, although in the same class, will attain distinct representations according to their scales in the feature space, which will directly benefit scale-aware density estima-

tions. We include a comprehensive comparison with various state-of-the-art methods for the task of crowd counting to verify the efficacy of incorporating geometric priors. Second, a multi-stage model with region-based supervisions is constructed to obtain robust features with implicit understandings of the scene geometry. With the internal multi-stage learning mechanism, features could be refined and adjusted repeatedly to perceive the scale variations. Besides, with local-based supervisions, the model is further constrained to generate locally consistent densities that conform to object scale variations. Experiments are presented to validate the effectiveness of the proposed model for crowd counting. Third, we build a multi-task framework that drives the network to embed desired semantic/geometric/numeric attributes to handle various type of challenges for crowd counting. With the multi-fold regularization effects introduced by three auxiliary tasks, the intermediate features are driven to convey desired properties and thus help improve the main task of density estimation. Extensive experiments have been conducted to indicate the effectiveness of the proposed method.